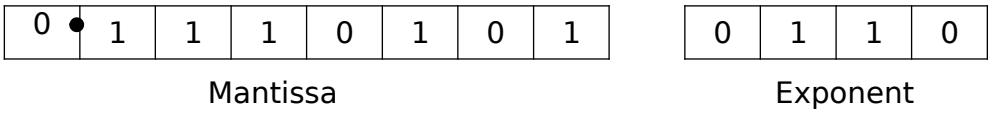
# Worksheet 4 Floating point arithmetic

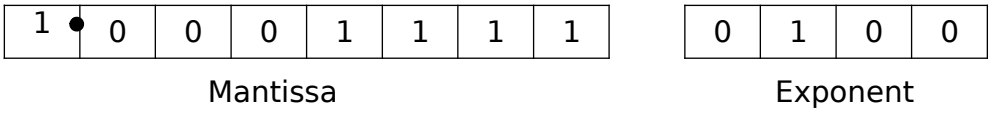
## Task 1

1. Convert the following floating point numbers from binary to decimal. Show your working.

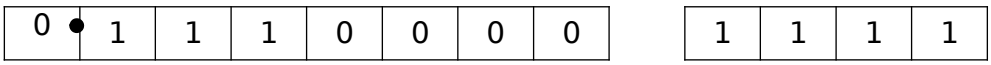
(a)



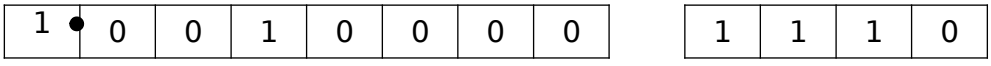

(b)




(c)




(d)

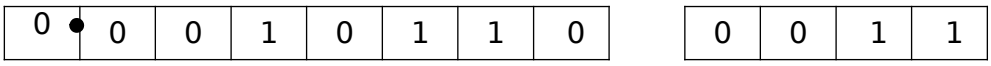


2. What is the largest number, in decimal that can be represented using this floating point system?

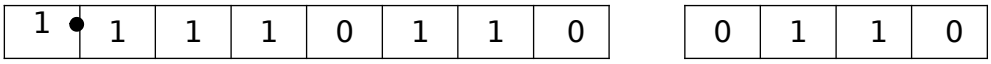
## Task 2

3. Convert the following binary numbers into normalised form:

(a)




(b)




4. Convert the following from decimal to normalised binary floating point, using an 8-bit mantissa and a four-bit exponent. show your working.

(a) 45.5

(b) -14.5

5. What is the most negative number that can be held in an 8-bit mantissa and a 4-bit exponent? Express the answer as a normalised floating point binary number.

## Task 3

6. Add together the two normalised binary numbers shown below, giving the result in normalised floating point binary form.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 • | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 • | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| • | | | | | | | | | | | |

7. Subtract the second binary number below from the first, giving the result in normalised floating point form.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 • | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 • | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

Convert the numbers to fixed point form

first number (A)

second number (B)

one's complement of B:

+ 1

two's complement (-B)

first number (A)

-B + A

Normalise by moving binary point n places

Add n to the exponent

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | • | | | | | | | | | | |

8. Subtract the second binary number below from the first, giving the result in normalised floating point form.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | • | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 0 | • | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| | • | | | | | | | | | | |